A Study of Image Colourfulness

Cristina Amati Niloy J. Mitra Tim Weyrich

University College London

Figure 1: *Which image is more colourful? Which image is more aesthetically pleasing? Such judgements are very obvious to humans but depend on complex processes that are very difficult to identify or describe. In this study we investigate the correlation between existing colourfulness metrics and data gathered from users but also if there is any link between colourfulness and image aesthetics. (Images by Flickr users Alexey Kljatov, Luca Argalia, Lutz Koch and Stefan Perneborg, respectively.)*

Abstract

Colourfulness is often thought of as a mere measure of quantity of colour, but user studies suggest that there are more factors influencing the perception of colourfulness. Boosting and enhancing colours are operations often performed for improving image aesthetics, but the relationship between colourfulness and aesthetics has not been thoroughly explored. By gathering perceptual data from a large-scale user study we have shown how existing colourfulness metrics relate to it and that there is no direct linear dependence between colourfulness and aesthetics but correlations arise for different image categories such as: "landscape", "abstract" or "macro".

CR Categories: H.1.2 [User/Machine Systems]: Human information processing— [I.2.10]: Vision and Scene Understanding— Perceptual reasoning

Keywords: colourfulness, aesthetics, crowd-sourced user study, perception

1 Introduction

Colour, as much as composition in an image, is a very important means of conveying messages and appealing to viewers' perceptual triggers. The physiology of the visual cortex highlights this importance by receiving colour stimuli in the pre-attentive stage as well as having a dedicated area for colour processing [Zeki 1999]. In Computer Graphics, colour is of great importance from image enhancement algorithms to appearance modelling and tone mapping, hence, to make our results more convincing, we need to base image assessment metrics such as colourfulness, colour harmonies, or contrast levels on perceptual data. Knowledge of how humans perceive and interpret has changed the way we approach problems in graphics from producing appealing visualisations to rendering based on human focus limitations, as acknowledged by [Bartz et al. 2008].

Colour has been studied in depth throughout the centuries. As a result, there are many detailed rules for colour harmonies in palettes isolated from actual images, but little effort has been expended towards creating a perceptually based model for a holistic interpretation of colour in an image, which we call *colourfulness*. (Note that this differs from the more low-level definition of perceptual colourfulness as used by Colour Appearance Models (CAMs) [Kuang et al. 2007] that measure the perceptual response to isolated colours on neutral grey backgrounds and do not take into account the combined effect of spatially distributed colours in an image.) Moreover, the relationship between amount and quality of colour and the aesthetic appeal of an image is not fully understood and scarcely explored. For example, most natural images have very limited colour palettes, yet can be perceived by people as being very colourful. High saturation and complementary colours account for such effects as much as a multitude of diverse colours.

In order to shed some light on the aforementioned problems, we ask ourselves three questions: (i) *Is there a consensus amongst people on colourfulness perception?* (ii) *How do existing colourfulness metrics correlate to perceptual data?* (iii) *Is there a relationship between colourfulness and aesthetics?*.

To answer those questions, we deployed two user studies, a pilot study to gather semantic data and a large-scale user study on colourfulness and aesthetics. The surveys were deployed via Amazon Mechanical Turk (MTurk) to a large heterogeneous pool of workers with great variation of taste, expertise, background and age. Our pilot study showed that according to user perception, attributes such as *vividness* are more important to colourfulness than *number of colours*. This gave us a good basis for the larger scale user study.

In our main survey, we tested pairwise comparisons on 100 images collected from Flickr. To our knowledge, this is the first attempt at assessing colourfulness through comparisons of random images as opposed to modified versions of the same image. This gives us a much more natural and realistic response. Another advantage of the pairwise approach is that an image ranking falls out naturally from the image pair ratings, without users having to worry about rating scales.

In total, the 100 images yield 4950 image pairs and to be able to establish user agreement the whole dataset has been evaluated 5

CAe 2014, August 08 – 10, 2014, Vancouver, British Columbia, Canada.

ACM 978-1-4503-3019-0/14/08

times. A total number of 516 users participated in our survey. Our image selection has been as neutral as possible, avoiding pictures of people and animals that might trigger a strong emotional response. The dataset is available for research use at the project webpage.

The data collected presented good user consensus on the image ratings. Moreover, we compared this data to two metrics from literature and our own colourfulness metric. We found that the perceptually based metric of Hasler and Süsstrunk [2003] was closest to the user responses, while the colourfulness metric most frequently used in literature, proposed by Datta et al. [2006] does not reflect perceptual data.

We found that aesthetics do not correlate directly with colourfulness because they are governed by more complex factors, including semantic ones such as image subject. On separating our dataset into categories such as "abstract", "landscape", "urban" and "macro" we discovered that correlations with colourfulness started emerging.

2 Related Work

Colour, when present, is one of the most salient features of an image. Many different fields have treated the issue of the nature of colour, colour perception and its application in creating images.

2.1 Colour Theories

Theories on colour started developing in Ancient Greece. Inspired by Aristotle, Leonardo developed his own set of primary colours and explained how they should be combined for maximal effect in paintings [Leonardo 1651]. Newton [1704] described the physical phenomenon behind colour and provided a geometrical arrangement based on wavelengths. Goethe [1810] defied Newton's theories and devised his own colour wheel based on physiological phenomena such as after-images, which led the way to complementary colours. One of the most influential colour theorists was the French chemist Chevreul. In his work [Chevreul 1839] he proposed the Law of Simultaneous Contrast which implies that the brain tends to exaggerate differences in hue and lightness to perceive them better. His theories were used by painters such as Delacroix, Signac and Monet to enhance colour appearance in their paintings. He was also the first to draw attention that gilded frames did not flatter paintings, which inspired the Impressionists to be the first to use frames that complemented the colours in their works.

2.2 Neuroaesthetics and Perception

Starting with Goethe and Chevreul, it was clear that there is a discrepancy between measurable properties of colour and our perception of these properties. The field of Neuroaesthetics aims at understanding what mental processes underlie the production and enjoyment of art. Zeki, who introduced Neuroaesthetics to the scientific community, believes that art cannot be studied in disjunction from the brain because the role of the artist is very similar to that of the visual brain: "a search for essence and constancy in an ever changing environment with the aim of obtaining information about the world" [Zeki 1999]. Another pioneer of the field, Ramachandran, proposes eight laws of artistic experience [Ramachandran and Hirstein 1999] and posits that most of our perceptions and reactions are rooted in human evolution. It is Ramachandran's pertinent belief that without the exploitation of hypernormal stimuli, there would be no art, only reproduction. Zeki also claims that accomplished artists have a masterful intuition of how the visual brain works.

Gestalt psychology is also concerned with pinning down the rules of visual perception. Laws such as Grouping, Common Fate, Closure, Symmetry and Past Experience, show how the brain tries to cope with new information by applying patterns and abstractions [Arnheim 1954].

2.3 Colour Harmonies

Art and graphical design rely heavily on colour to convey messages. Although there are no definite recipes for how colours should be put together quantitatively and qualitatively, several models for colour harmonisation exist. Such patterns have been developed by Munsell [1921], Moon and Spencer [1944], Itten [1973] and Matsuda [1995] based on psychophysical data or geometrical laws. All these works concern patterns of 2 to 5 colours and are heavily used in colour harmony related research in Computer Graphics.

Ou et al. [2004] study colour emotion and preference for single colours and two colour combinations in which they find colour preference is dictated by subjective factors such as personal taste and cultural background and a large number of people dislike colour combinations that they considered harmonious. Schloss and Palmer [2010] perform a detailed study of colour pair harmony using controlled environment user studies.

Burchett [2002] identifies the attributes that influence colour harmony: Order, Tone, Configuration, Interaction, Similarity, Association, Attitude, and Area. Association and Attitude are purely subjective factors that depend on the viewer's mood and background whilst the others adhere to principles described before in literature of reciprocal influence of juxtaposed colours.

Automatic colour harmonisation in an image has been achieved by Cohen-Or et al. [2006], by optimising the image hue histogram to fit the Matsuda colour templates. Results are impressive, but they do not take into account how hue manipulation affects the colourfulness or contrast of the image. O'Donovan et al. [2011] look at creating harmonious colour templates by harnessing data from online communities and Amazon Mechanical Turk, while Lin and Hanrahan [2013] model the way people extract colour themes from images.

Relating colour harmonies to aesthetics, Nishiyama et al. [2011] use the Moon-Spencer and Matsuda models to extract local pairwise colour harmonies from images and use them in a bag-of-words inspired approach for aesthetics quality classification.

Heer and Stone [2012] cross over to the semantic domain and develop a probabilistic colour naming model and demonstrate how this approach greatly improves tasks such as name-based pixel selection methods for image editing, and evaluating colour palette design.

2.4 Colourfulness Metrics and Aesthetics

Metrics for colourfulness have been proposed in the context of image compression quality evaluation by Hasler and Süsstrunk [2003] and in the context of aesthetics inference by Datta et al. [2006].

The metric proposed by Datta et al. represents a candidate feature for a machine learning framework that they build for aesthetics classification of images. Although this metric was never perceptually validated or claimed to be so, it had been used in further image classification related literature (e.g. [Machajdik and Hanbury 2010]).

Hasler and Süsstrunk's metric is fitted to perceptual data collected from a controlled user study and it has a simple expression (detailed in 6.2) based on the a*b* pointcloud of the image in L*a*b* colour space. San Pedro and Siersdorfer [2009] use this colourfulness metric amongst other features for ranking and classifying photographs on the web according to attractiveness.

3 Experimental Design

In light of many studies on colour and perception it is clear that there are some strong tendencies and preferences but no definite answers to colour harmonies and colourfulness. We know that our intuitive perception of colour does not always correlate with the properties that we measure because often we fail to take into account the complex interactions between the colours themselves.

In order to test our hypothesis that there is a consensus amongst users regarding colourfulness, we have deployed a pilot user study and then a larger scale colourfulness study using Amazon Mechanical Turk (MTurk). This provides a convenient way to reach a large heterogeneous pool of people from all around the world. As both studies were deployed via web platforms, they have been carried out in uncontrolled conditions, hence a consensus amongst users would be even more difficult to reach but all the more meaningful. The viability of using MTurk for visual tasks has been studied and found viable by Heer and Bostock [2010].

Both user studies were built as pairwise image comparison studies because giving an absolute score on a certain scale is a difficult and error-prone task and people find it much easier and natural to compare two objects. Thus, in order to find a ranking of images by colourfulness we have generated all the possible pairwise comparisons between the images in our dataset and had users evaluate them. A perceptually based ordering of the images, then, falls naturally out of the pairwise comparison data.

3.1 Pilot Study

First off, we designed a pilot study in which we wanted to test the pairwise comparison method and collect free text data about people's perception of colourfulness.

This study was carried out on a restricted set of 20 images collected from Flickr and consisted of three tasks. In the first task, the image pairs were presented side by side on a neutral gray background and users were asked to choose the one they considered more colourful. This was not a forced choice comparison and "equally colourful" was allowed as an answer. The second task presented a single image to the user and asked for an absolute rating of colourfulness on a scale of 1 to 4, descriptions of the salient colours in the image and which colours they found contrasting. In the final task, whilst re-iterating the pairs from Task 1, users were asked to re-evaluate them in order to test consistency. They were also asked to explain their choice so that we could identify the main attributes of colourfulness.

3.2 User Study

Our second study was carried out on a much larger scale, using 100 different images which resulted in 4950 pairwise comparisons. Although a considerable amount of data, we have opted for a complete rather than incomplete pairwise comparison method to be able to mitigate the noise in the user data. Furthermore, we collected 5 redundant passes of the whole dataset for purposes of establishing user agreement and outlier filtering. This amounted to a total of 24750 pairwise comparisons.

Each user was presented with 20 pairs of images and two control image pairs. This would constitute a Human Intelligence Task (HIT) on MTurk and workers were asked to complete no more than 3 such HITs so that we could collect as many opinions as possible. In total, we needed 1240 HITs in order to complete our study. To avoid bias, the image pairs for each HIT were randomly assigned and each redundant pass of the dataset was separately randomly generated thus making all of the 1240 HITs unique.

For each image pair, workers were asked 4 multiple-choice questions: Q1) "Which image looks more colourful?" Q2) "How confident are you of your response?" (regarding colourfulness) Q3) "Which image looks more aesthetically pleasing?" Q4) "How confident are you of your response?" (regarding aesthetics). Multiple-choice answers for Q1) and Q3) were: a) "Right image", b) "Equally colourful/pleasing" c) "Left image". The buttons corresponding to these answers were placed so that they correlate visually with the image positions (see Figure 2). The confidence-related answers Q2) and Q4) were ordered on a Likert scale of 4 with equal distance among: a) "Not confident", b) "Slightly confident", c) "Moderately confident", d) "Very confident".

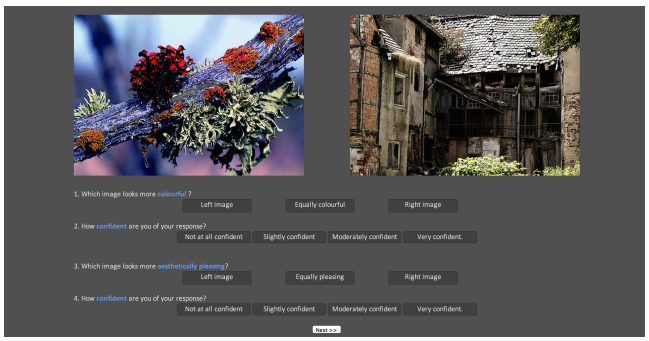


Figure 2: *Task layout for the large-scale user study.*

3.3 Image Selection

The images used in this study have been selected from the online photo sharing community Flickr and we specifically chose images under Creative Commons License.

The subject of the image is very important in triggering affective or repulsive reactions, which might skew the aesthetics judgement. To keep such distractions to a minimum, we chose neutral images from four categories: landscape, macro, urban and abstract. The abstract category includes geometrical compositions (Figures 4a, 4b) and we consider a macro a picture with a clear subject of focus (Figures 4e, 4f). We tried as much as possible not to include people, animals or objects that might trigger a strong emotional response. Thus, our image selection process was carried out using keyword searches on Flickr for terms such as "colourful", "landscape", "objects", "cityscape", "abstract" and combinations thereof. For each category we downloaded manually a series of candidate images with appropriate subject and then further refined the selection based on colourfulness.

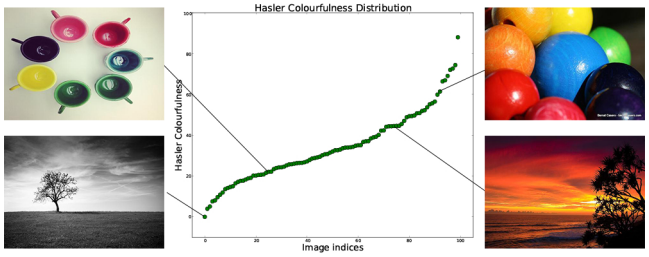
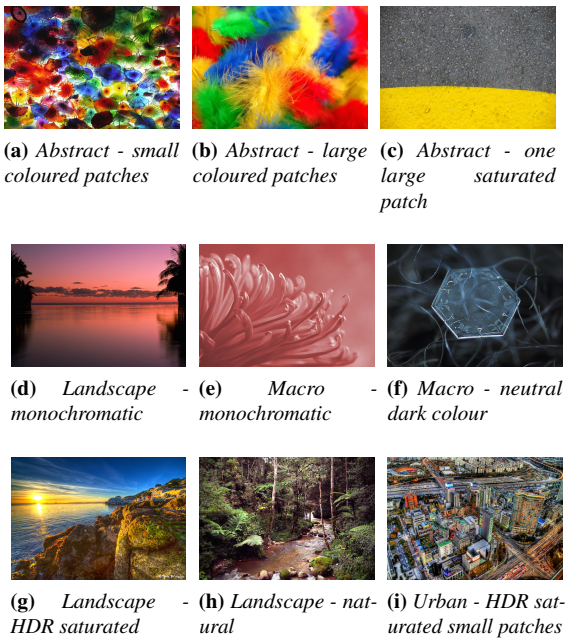


Figure 3: *Distribution of colourfulness in user study images using Hasler and Süsstrunk's colourfulness metric as a heuristic. Images by Flickr users: Shapeways (cups), Tobi Gaulke (tree), Bernat Casero (spheres) and Michael Dawes (sunset).*

In order to ensure an even distribution of colourfulness in our dataset, we used Hasler and Süsstrunk's colourfulness metric as a heuristic.



(a) *Abstract - small coloured patches* **(b)** *Abstract - large coloured patches* **(c)** *Abstract - one large saturated patch*

(d) *Landscape - monochromatic* **(e)** *Macro - monochromatic* **(f)** *Macro - neutral dark colour*

(g) *Landscape - HDR saturated* **(h)** *Landscape - natural* **(i)** *Urban - HDR saturated small patches*

Figure 4: *Selection of images from the large-scale user study. Images by Flickr users Slices of Light, Lucy Nieto, Lali Masriera, Pedro Szekely, casch52, Alexey Kljatov, David Yu, Zaqqy and Trey Ratcliff, respectively.*

This distribution is shown in Figure 3. Our selection included black and white images, sepia toned images, as well as very colourful images of two varieties: 1. many colours in small patches and (Figure 4i) 2. fewer colours in large saturated patches (Figure 4b).

Figure 4 shows a selection of images from our dataset pertaining to different image categories and having different properties related to colour patch size, saturation and number of colours. We have chosen pairs that we deemed difficult to asses, such as Figures 4a and 4b with same levels of saturation but different patch sizes, Figures 4c and 4d with similar palette but different contrast levels.

4 Data Collection and Cleaning

The user study described above produced five user responses for each of the possible pairwise comparisons between images.

4.1 Demographics

Our study was performed by a total of 516 users, aged 18 to 73 with 63% male and 37% female participants. We have also asked them to state their experience with visual arts, photomanipulation and computer graphics; we discovered a heterogeneous distribution of experience across all these criteria. The expert users in each of these domains were the fewest: 7% Photoshop experts, 6.37% Computer Graphics experts and 5.8% highly skilled artists. There were between 19% and 28% of users inexperienced in any of these fields, whilst around 40% of users were beginners and around 30% had intermediate skills.

4.2 Culling Insincere Responses

Although a very convenient medium to reach thousands of users, Mechanical Turk does not guarantee the quality of its workers. To safeguard against dishonest workers we have logged each button click and set up two control questions with obvious answers. The images for Control Pair 1 can be seen in Figure 5.

To filter out random clickers we compute a confidence score. There are eight criteria that are likely to characterise a random clicker: 1) the answer to either of the Control Pairs is wrong, 2) the answer to Control Pair 1 is wrong - we award an extra point if such an obvious comparison is incorrectly evaluated, 3) the average time taken to answer each question is under a second - this is a sign of a hasty user, 4) the questions for at least one image pair are answered out of order starting with the confidence level at the bottom of the page, 5-8) the distribution of neutral answers (situated in the middle of the page) is higher than 60% for each answer.

Another culling criterion is based on a request made to workers to not take more than 3 jobs from the whole study. Users that have taken more than 6 turned out to be 90% random clickers, including one user who went through 112 jobs by randomly clicking on buttons.

In total we discarded 242 HITs out of the available 1240 (almost 20%) due to random clicking and workers exceeding their requested allotment of jobs. The jobs were reposted until completed in a satisfactory manner.

(a) *Control Pair 1a* **(b)** *Control Pair 1b*

(c) *Control Pair 2a* **(d)** *Control Pair 2b*

Figure 5: *Control Pairs 1 and 2. Control Pair 1 is a very obvious choice, hence evaluating it incorrectly is a sign of a possible insincere user. (Images by Flickr users: Lali Masriera, Michele Catania, Des Wass and Stewart Baird, respectively.)*

4.3 Handling Noise in MTurk Responses

Such an amount of data processed by a large number of people is prone to noise. Even though we have taken precautions and detected many insincere users, it is virtually impossible to guess all the ways in which users might be dishonest in completing this task.

In a cumulative analysis of each pass of the 4950 comparison dataset, we used the Kendal-Tau metric (see Section 5.4) to compare the image rankings produced by 1, 2, 3, 4 and then all the evaluations of the comparison dataset. We can see from Table 1 that the difference between 4 and 5 passes is reduced in comparison to the difference between 3 and 4 passes or 1 and 2 passes. This shows that given enough passes of the dataset, the image rankings will eventually converge to a stable position.

The differences between individual passes are in the range $[0.069, 0.097]$ which shows that very small changes occur between the ratings of distinct user groups.

5 Methodology

We describe the tools we used in our data analysis.

5.1 User Agreement Computation

As there were no two identical HITs, each user was confronted with a different set of decisions, hence usual methods of computing user consensus do not apply. Our method of computing user agreement is based on the fact that we have gathered 5 redundant evaluations of each image pair. If three or more people agree on the rating of that pair, we consider that to indicate consensus. Results on user agreement will be reported in Section 6.1 for colourfulness and in Section 6.3 for aesthetics.

5.2 Image Ranking from Pairwise Comparisons

Once we have retrieved all image comparisons for our dataset we proceed to order the images by colourfulness. We do so by computing a score for each image based on how many times it was chosen over the other images. If the total number of images is N, then each image appears $N-1$ times within the full set of $N(N-1)/2$ pairwise comparisons. The score for image k is:

$$s_k = \sum_{i=1}^{N-1} p_{ki}, \tag{1}$$

where $p_{ki} \in \{0, 0.5, 1\}$ is the amount awarded to image k when compared to image i. p_{ki} can take one of three values: 1 if image k was chosen over image i, 0.5 if there was a tie and 0 if image i was chosen over image k. Thus, the most colourful image will have been chosen most times over the other images and obtained the highest score. It can also happen that some images obtain the same score in which case we do not try to forcedly break the tie as it would alter the user data.

5.3 Confidence Weighted Image Rankings

Some image pairs are more difficult to judge than others. For this purpose we have asked users for the confidence of their response. The confidence was given on a Likert scale of 1 to 4 as follows: 1) not confident, 2) slightly confident, 3) moderately confident, 4) very confident. We have chosen an even scale to avoid the tendency of users to choose the neutral middle value. We recompute the scores for each image similarly to the method of Section 5.2 but this time we weight the awarded point by the confidence value. The confidence value set is $\mathrm{Conf} = \{1, 2, 3, 4\}$. Because we chose a Likert scale, the distance between confidence points is considered equal and for our 4 point scale it amounts to 0.25. The weighted score for one image is:

$$sw_k = \frac{\sum_{i=1}^{N-1} c_{ki} p_{ki}}{\sum_{i=1}^{N-1} c_{ki}}, \tag{2}$$

Table 1: *Kendall-Tau rank correlation between image rankings obtained using increasing number of redundant passes of the comparison dataset. We can see by looking at the first diagonal that the difference between progressive numbers of passes decreases as the number of passes increases.*

Number of datasets	1	2	3	4	5
1	-	0.036	0.042	0.043	0.043
2	-	-	0.026	0.034	0.037
3	-	-	-	0.019	0.022
4	-	-	-	-	0.013

where $p_{ki} \in \{0, 0.5, 1\}, c_{ki} \in \mathrm{Conf}$.

5.4 Comparing Rankings

In order to compare perceptual data to existing colourfulness metrics, we rank the 100 images in our dataset using all these methods and then compare the resulting rankings using the Kendall-Tau metric [Kendall 1938]. The Kendall-Tau rank correlation computes the percentage of image pairs that are ranked differently by the two methods being compared. For two rankings τ_1 and τ_2, $K(\tau_1, \tau_2) = ||\{(i,j) : i < j, ((\tau_1(i) < \tau_1(j)) \wedge (\tau_2(i) > \tau_2(j))) \vee ((\tau_1(i) > \tau_1(j)) \wedge (\tau_2(i) < \tau_2(j)))\}||$.

For a set of N images we will have $M = N(N-1)/2$ possible pairwise combinations. We normalise $K(\tau_1, \tau_2)$ by M to obtain a number in the interval $[0, 1]$, where 0 means total correlation and 1 means total discrepancy between the two rankings being compared.

5.5 Measuring Linear Correlation

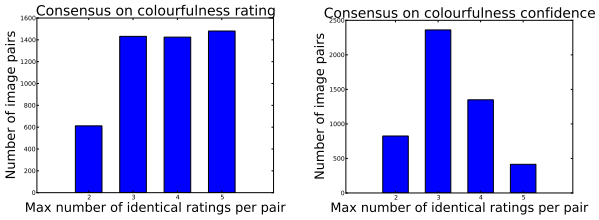
For measuring correlation between aesthetics and colourfulness we use the Pearson's r product-moment correlation coefficient [Pearson 1895]. This gives an estimation of the linear correlation between two variables. The values are within the interval $[-1, 1]$, with 1 being total correlation, 0 lack of correlation and -1 total negative correlation.

6 Results

In this section we test our posited hypotheses and report on results from user data.

6.1 Is There User Agreement on Colourfulness?

Regarding our first hypothesis that there is a consensus on colourfulness rating amongst users, we have found that people agree on 87% of the image comparisons. We computed user agreement as described in Section 5.1. Figure 6a shows the distribution of the maximum number of identical ratings for all image pairs. The mean and standard deviation of the maximum number of identical ratings per pair are 3.76 and 1.01, respectively.



(a) *Rating agreement* **(b)** *Confidence agreement*

Figure 6: *Colourfulness rating and confidence agreement.*

To refute the null hypothesis that these observations could have arisen from a random distribution of image ratings, we have created 5 random ratings for each image pair in our dataset and used the same method as for human participants to compute agreement. For this random distribution we found agreement for 62.18% of the image pairs which is significantly lower than the 87% agreed by users. The mean and standard deviation of maximum number of identical ratings per par for the chance distribution are 2.77 and 0.70, respectively.

Confidence levels were given for each pair. Figure 6b shows the user agreement on confidence levels computed the same way as for

colourfulness. We see that there is overall agreement on 83.33% of the image pairs, although not as strong as for colourfulness rating. Mean and standard deviation for maximum number of identical ratings are 3.27 and 0.83.

For the scenario of random confidence rating, we have agreement on 42.18% image pairs, with a mean of 2.48 and standard deviation of 0.62.

Regarding the difficult image pairs that we described in Section 3.3, the images in Figures 4a and 4b were indeed difficult to assess with low consensus on the rating. Figures 4c and 4d did not pose a problem, with high agreement that 4c is more colourful than 4d.

6.2 Do Existing Colourfulness Metrics Correlate with User Data?

After having established user consensus we proceed to obtain a ranking of the images in our dataset according to colourfulness. We obtain 5 different rankings of the same image dataset using the methods described below:

1. User data pairwise comparisons (UD)
2. User data pairwise comparisons weighted by confidence values (UDW)
3. Datta et al. [2006] colourfulness metric (C_{Datta})
4. Hasler and Süsstrunk [2003] colourfulness metrics (C_{Hasler1} and C_{Hasler2})
5. Our contrast-based colourfulness metric (C_{Contr}).

From the pairwise comparisons rated by users we employ the method described in Sections 5.2 and 5.3 to rank the images. For colourfulness metrics, we compute the values for each image and then order them accordingly.

C_{Datta} is obtained by first dividing the RGB cube into 64 equal partitions and computing the frequency of colour within each partition. This distribution is then compared to an ideal distribution (D_1) of a colourful image where each RGB partition has a frequency of 1/64 by using the Earth Mover's Distance (EMD) [Rubner et al. 2000]. The pairwise distance metric needed for the EMD is that between the geometrical centers of each RGB subcube transformed to CIELUV colour space. Hence, their colourfulness metric is: $C_{\mathrm{Datta}} = \mathrm{EMD}(D_1, D_2, d(a,b)|0 \leq a,b \leq 63), d(a,b) = ||rgb2luv(c_a) - rgb2luv(c_b)||$.

C_{Hasler1} and C_{Hasler2} have been fitted to user data and computed in L*a*b* colour space. $C_{\mathrm{Hasler1}} = \sigma_{ab} + 0.37\mu_{ab}; C_{\mathrm{Hasler2}} = \sigma_{ab} + 0.94\mu_C$, where $\sigma_{ab} = \sqrt{\sigma_a^2 + \sigma_b^2}$ is the trigonometric length of the standard deviation in a*b* space, μ_{ab} is the distance of the centre of gravity in a*b* space to the neutral axis and μ_C is the mean Chroma.

We propose our own metric (C_{Contr}) based on colour contrast. For this, we cluster the image pixels in L*a*b* space using k-means clustering. For each cluster we obtain the cluster area normalised by the total image area a_k, the mean saturation of the cluster ms_k and the mean Euclidean distance to all the other cluster centroids, md_k. md_k will give us a measure of colour contrast. Thus, for N clusters, $C_{\mathrm{Contr}} = \sum_{i=1}^{N} md_k ms_k a_k$.

We have conducted a comparison between the colourfulness metrics described in Section 2.4, our contrast-based colourfulness metric and the perceptual data collected from users. We use the Kendall-Tau [Kendall 1938] distance described in Section 5.4 to compute the discrepancy between the image rankings produced by the various metrics. Table 2 shows the results.

Table 2: *Kendall-Tau rank correlation between image rankings obtained using the perceptual user data (UD), user data weighted by confidence (UDW), Hasler and Süsstrunk colourfulness metric (C_{Hasler1}) , Datta et al.'s EMD based colourfulness metric (C_{Datta}) and our contrast-based colourfulness metric (C_{Contr}).*

Method	UD	UDW	C_{Hasler1}	C_{Datta}	C_{Contr}
UD	-	0.010	0.187	0.436	0.247
UDW	-	-	0.189	0.436	0.248
C_{Hasler1}	-	-	-	0.430	0.179
C_{Datta}	-	-	-	-	0.443

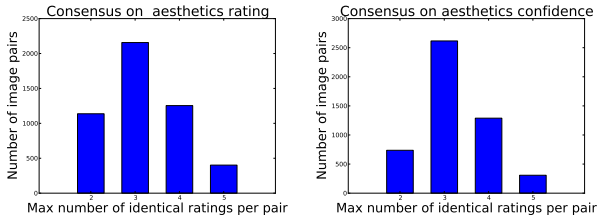
We can see that the perceptually based metric of Hasler and Süsstrunk is the closest to our experimental data whilst the frequency-based metric proposed by Datta does not correlate well with user responses. Our metric performs better than C_{Datta}, but not as well as C_{Hasler1} as it is semantically inspired by user interpretation of colourfulness but not directly fitted to user data.

6.3 Is There User Agreement on Aesthetics?

In our study we also asked users to rate image pairs according to aesthetic appeal in an attempt to learn whether colourfulness correlates at all with the beauty of an image.

After gathering results, we first wanted to see if there is user agreement with respect to aesthetics rating and confidence levels. The consensus was computed as described in Section 5.1 and Figure 7 shows that there is agreement on 85.11% of image pairs in terms of rating and 77.03% agreement on the confidence of these responses.

In the case of completely random answers to all aesthetics questions we have rating agreement on 64.16% of image pairs and for confidence levels on 41.13% of image pairs.



(a) *Rating agreement* **(b)** *Confidence agreement*

Figure 7: *Aesthetics rating and confidence agreement.*

We can see from Figures 6 and 7 that aesthetics agreement is not as strong as colourfulness agreement which shows that tastes definitely differ, but there is a baseline to aesthetics preferences. The mean of the maximum identical rating per image pair for aesthetics rating is 3.23 compared to a chance mean of 2.79. For confidence levels we have a user mean of 3.18 and a chance mean of 2.47.

6.4 Is There Any Correlation Between Colourfulness and Aesthetics?

Many image enhancing operations also attempt to saturate or harmonise colours. Individual colour palettes have been intensively studied but a holistic view of colour in an image was never thoroughly correlated with aesthetic appeal.

After computing colourfulness and aesthetics scores for all images as shown in Section 5.2, we have plotted colourfulness against aesthetics for all images. We use Pearson's r correlation coefficient

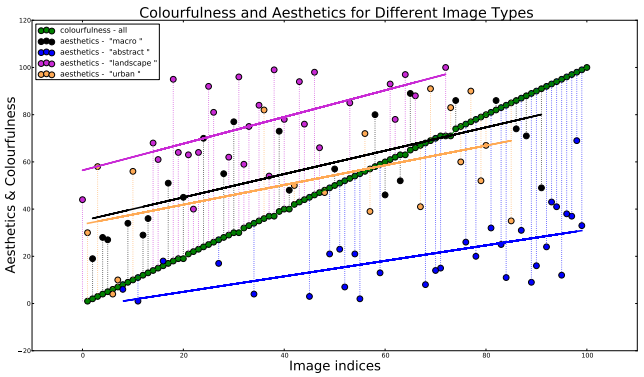


Figure 8: *Colourfulness vs. aesthetics for different image types: "landscape", "abstract", "macro" and "urban".*

described in Section 5.5 to measure the linear correlation between colourfulness and aesthetics. The result is a coefficient of 0.00569 with a p-value of 0.95 which shows that there is no correlation between the two and there is a high probability that an uncorrelated system could have produced such data.

On a first inspection of the image ranking according to aesthetics we have discovered that the top-ranked images were natural and landscape images. We divided up our image dataset into the four categories mentioned in Section 3.3: "landscape", "abstract", "macro" and "urban". If we order all images according to colourfulness and then plot the corresponding aesthetics scores for each image, we obtain Figure 8. We can see a clear clustering of the abstract images as least aesthetically pleasing, regardless of colourfulness values.

Table 3 shows the Pearson's r correlation coefficient between colourfulness and aesthetics for each image category. We can see that there is higher than chance correlation with statistically significant p-values, smaller than 0.01 for "landscape", "abstract", "macro" and slightly higher, 0.03 for "urban", which is the most scattered distribution.

Table 3: *Pearson's r correlation between colourfulness and image groups: "landscape", "abstract", "macro" and "urban".*

Image group	**Pearson's r**	**p-value**
Macro	0.699	0.0002
Abstract	0.585	0.0005
Landscape	0.584	0.0010
Urban	0.502	0.0335
All	0.005	0.9551

7 Predicting Colourfulness Ranking

Having collected colourfulness data from users, we can try to predict the colourfulness of other images. As we are dealing with pairwise comparisons as input, we use SVMRank [Joachims 2006] to train an SVM to perform pairwise comparisons similarly to the users in our study. The features we use for training the SVMRank are L*a*b* colour histograms with 7 bins per channel. To apply the learned model to rank a set of new images, we first generate all pairwise comparisons between these images, predict the rating for each comparison using the SVMRank model and then rank the image using the method in Section 5.2.

To test the predicted ranking, we compare it against rankings produced by the metrics described in Section 6.2 but also against more user data collected from two small user studies (AMT_1 and AMT_2) conducted via MTurk and set up similarly to our large-scale study. All rankings were computed on the same set of 10 landscape images with various degrees of colourfulness. We have chosen landscapes to avoid subject bias.

Table 4: *Kendall-Tau rank correlation on a 10 image dataset between image rankings obtained using SVMRank predictions, the 3 metrics and two small scale MTurk user studies.*

Method	SVMRank	C_{Hasler1}	C_{Datta}	C_{Contr}	AMT_1
C_{Hasler1}	0.044	-	-	-	-
C_{Datta}	0.6	0.6	-	-	-
C_{Contr}	0.155	0.2	0.577	-	-
AMT_1	0.177	0.222	0.555	0.2441	-
AMT_2	0.244	0.288	0.488	0.177	0.2

The rankings were compared using the Kendall-Tau rank correlation and results can be seen in Table 4. The images and their ranking by each method is presented in Figure 9.

7.1 Discussion

We can see from Table 4 that SVMRank predictions align well with the new user data and also that the two groups of users from AMT_1 and AMT_2 have divergent opinions on the colourfulness of the given image dataset. The two user groups agree on the five most colourful and five least colourful images, but within these two categories there is little consensus. It can be seen that people consider as colourful images that have at least two contrasting colours of moderate to high saturation and lightness or a multitude of differently coloured small patches. The least colourful image is agreed by both groups to be the top image in the first column of Figure 9.

It is encouraging to see that the SVMRank predictions learned from the data in our large-scale user study correlates well with the results of AMT_2 and AMT_1 in particular. Our contrast based colourfulness metric is closest to the preferences expressed in AMT_2 and outperforms all other colourfulness computation for this particular case. The discrepancy between the two user groups is a reminder of the subjectivity of the colourfulness judgement and the fact that we can only model perceptual approximations for groups of people - the larger the group, the more inclusive the approximation.

One observation is constant throughout all data: the fact that any perceptually based metric is much more accurate than non-perceptual ones. C_{Hasler1} and C_{Contr} consistently outperform C_{Datta} on all datasets. Our metric is not fitted directly to user data but it is semantically inspired by the results of our pilot study and our large-scale user study.

8 Conclusions and Future Work

In this work we have shown that basing metrics for colourfulness on perceptual data is beneficial for their accuracy and descriptiveness. We conclude that large-scale user studies, although instrumental in getting perceptual insight will only model the preferences of a group of people more or less broadly.

Following up on the textual information from our pilot study and the conclusions from the small follow-up user studies, we think that there is value in trying to semantically decompose colourfulness into attributes such as "contrast", "saturation", "area size" etc., attributes that have also been described by cognitive psychology work [Burchett 2002]. Much attributes related work has emerged from the Machine Vision and the Machine Learning community and

it proves that such mid-level layers are useful in modelling complex human judgements [Parikh and Grauman 2011].

Finally, we want to learn more about the relationship between use of colour and aesthetics for different types of images and compositions. As we have seen in our data, individual image groups have different rules with respect to colour usage and what might be aesthetic for abstract images might be unsightly for a landscape.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback and suggestions. We are also grateful to David Hasler and Sabine Süsstrunk for making their colourfulness study data available, to Craig Kaplan for helpful comments and James McRae for insight with MTurk. This work has been supported by the EngD VEIV Centre for Doctoral Training, Anthropics Technology Ltd., the Marie Curie Career Integration Grant 303541 and the ERC Starting Grant SmartGeometry (StG-2013-335373).

References

ARNHEIM, R. 1954. *Art and Visual Perception*. University of California Press, Berkeley and Los Angeles.

BARTZ, D., CUNNINGHAM, D. W., FISCHER, J., AND WALLRAVEN, C. 2008. *State-of-the-Art of the Role of Perception for Computer Graphics*. Blackwell, 65–86.

BURCHETT, K. E. 2002. Color harmony. *Color Research & Application 27*, 1, 28–31.

CHEVREUL, M. E. 1839. The principles of harmony and contrast of colours and their application to the arts.

COHEN-OR, D., SORKINE, O., GAL, R., LEYVAND, T., AND XU, Y.-Q. 2006. Color harmonization. *ACM Trans. Graph. (Proc. SIGGRAPH) 25*, 3 (July), 624–630.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part III*, Springer-Verlag, Berlin, Heidelberg, 288–301.

GOETHE, J. W. 1810. *Theory of Colors*.

HASLER, D., AND SÜSSTRUNK, S. 2003. Measuring Colourfulness in Natural Images. In *Proc. IS&T/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, vol. 5007, 87–95.

HEER, J., AND BOSTOCK, M. 2010. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM Human Factors in Computing Systems (CHI)*, 203–212.

HEER, J., AND STONE, M. 2012. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, CHI '12, 1007–1016.

ITTEN, J. 1973. *The Art of Color: the subjective experience and objective rationale of color*. Van Nostrand Reinhold, New York.

JOACHIMS, T. 2006. Training linear SVMs in linear time. In *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD)*, 217–226.

KENDALL, M. 1938. A new measure of rank correlation. In *Biometrika*, 81–89.

KUANG, J., JOHNSON, G. M., AND FAIRCHILD, M. D. 2007. iCAM06: A refined image appearance model for HDR image rendering. *Journal of Visual Communication and Image Representation 18*, 5 (Oct.), 406–414.

LEONARDO. 1651. On colours. In *Leonardo on Painting*, M. Kemp and M. Walker, Eds. 70–76, Yale University Press, 2001.

LIN, S., AND HANRAHAN, P. 2013. Modeling how people extract color themes from images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 3101–3110.

MACHAJDIK, J., AND HANBURY, A. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the International Conference on Multimedia*, ACM, 83–92.

MATSUDA, Y. 1995. *Color Design*. Asakura Shoten.

MOON, P., AND SPENCER, D. E. 1944. Geometric formulation of classical color harmony. *J. Opt. Soc. Am. 34*, 1 (Jan.), 46–50.

MUNSEL, A. 1921. *A Grammar of Color*.

NEWTON, I. 1704. *Opticks or, a treatise of the reflexions, refractions, inflexions and colours of light : also two treatises of the species and magnitude of curvilinear figures.*

NISHIYAMA, M., OKABE, T., SATO, I., AND SATO, Y. 2011. Aesthetic quality classification of photographs based on color harmony. In *Computer Vision and Pattern Recognition*, IEEE, 33–40.

O'DONOVAN, P., AGARWALA, A., AND HERTZMANN, A. 2011. Color compatibility from large datasets. *ACM Trans. Graph. 30*, 4 (July), 63:1–63:12.

OU, L.-C., LUO, M. R., WOODCOCK, A., AND WRIGHT, A. 2004. A study of colour emotion and colour preference. part III: Colour preference modeling. *Color Research & Application 29*, 5 (Oct.), 381–389.

PARIKH, D., AND GRAUMAN, K. 2011. Interactively building a discriminative vocabulary of nameable attributes. In *Computer Vision and Pattern Recognition*, 1681–1688.

PEARSON, K. 1895. Notes on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*, vol. 58, 240–242.

RAMACHANDRAN, V. S., AND HIRSTEIN, W. 1999. The science of art: A neurological theory of aesthetic experience. *Journal of Consciousness Studies 6*, 6–7, 15–51.

RUBNER, Y., TOMASI, C., AND GUIBAS, L. J. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision 40*, 2 (Nov.), 99–121.

SAN PEDRO, J., AND SIERSDORFER, S. 2009. Ranking and classifying attractiveness of photos in folksonomies. In *Proc. of the 18th International Conf. on World Wide Web*, ACM, 771–780.

SCHLOSS, K. B., AND PALMER, S. E. 2010. Aesthetics of color combinations. In *Human Vision and Electronic Imaging*, SPIE, vol. 7527 of *SPIE Proceedings*, 752719.

ZEKI, S. 1999. Art and the brain. *Journal of Consciousness Studies 6*, 7 (June), 76–97.

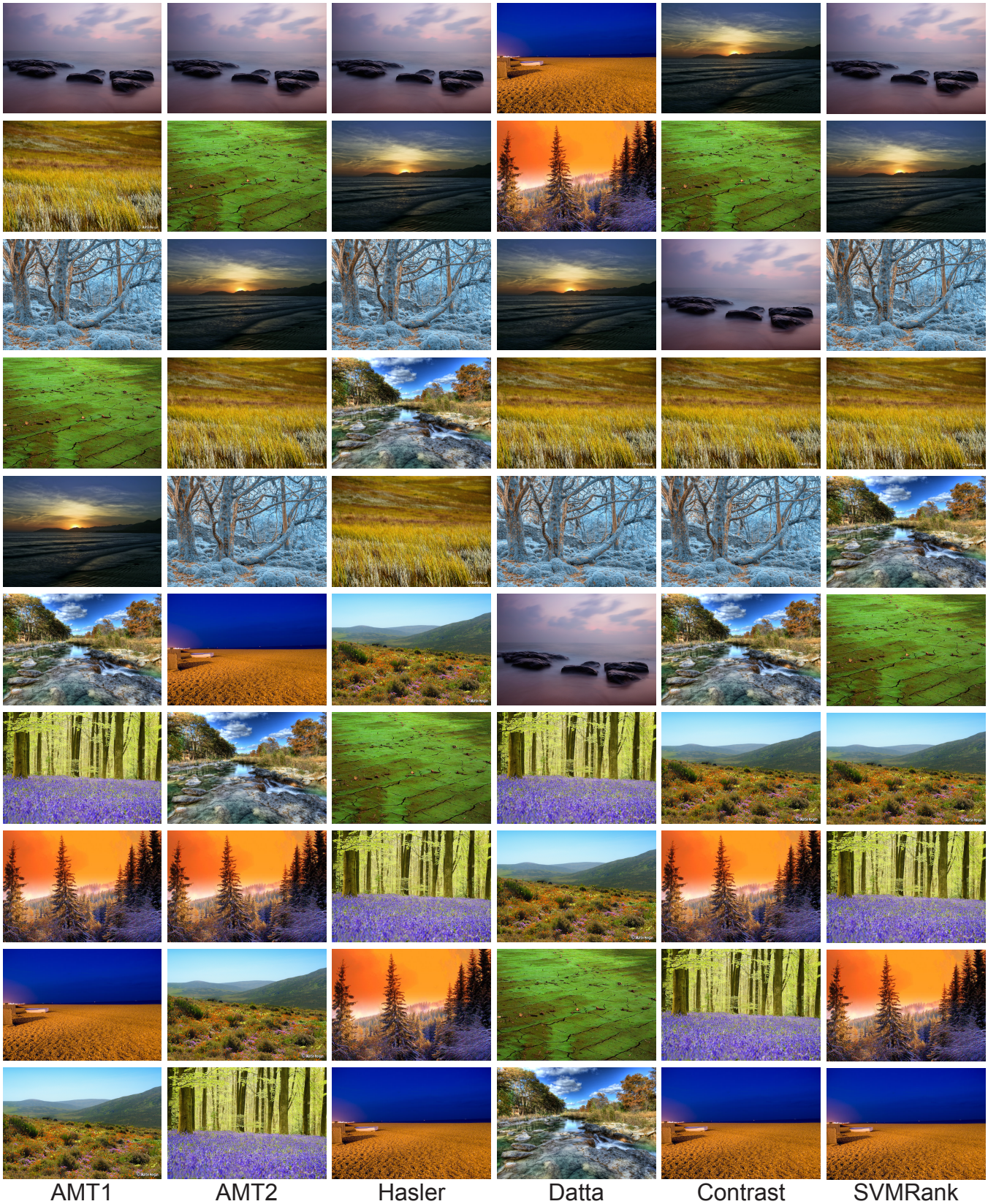


Figure 9: *Image rankings produced by various methods: AMT1 and AMT2 resulted from two MTurk user studies, Hasler, Datta and Contrast were obtained with the following colourfulness metrics: C_{Hasler1}, C_{Datta} and C_{Contr} respectively. SVMRank ranking was produced from the predictions made using data from our large-scale user study. Images are arranged from top - least colourful to bottom - most colourful. (Images following the first column top to bottom, by Flickr users: Vinoth Chandar, Martin Heigan, Nicolas Raymond, Hejma, Jeremy Raff-Reynolds, Mark Schaffer, Tony Braime, Stella Momcheva, wagdi.co.uk and Martin Heigan.)*